6.896 Sublinear Time Algorithms	February 27, 2007
Homework 2	
Lecturer: Ronitt Rubinfeld	*Due Date: March 15, 2007*

Homework guidelines: You may work with other students, as long as (1) they have not yet solved the problem, (2) you write down the names of all other students with which you discussed the problem, and (3) you write up the solution on your own. No points will be deducted, no matter how many people you talk to, as long as you are honest. If you already knew the answer to one of the problems (call these "famous" problems), then let me know that in your solution writeup – it will not affect your score, but will help me in the future. It's ok to look up famous sums and inequalities that help you to solve the problem, but don't look up an entire solution.

Turn in your solution to each *problem on a separate sheet of paper, with your name on each one.*

1. (Lower bounds)
   (a) Show that there are two distributions a and b such that $||a-b||_2 = \epsilon$ and any algorithm that is given $o(1/\epsilon^2)$ samples from each of two distributions cannot distinguish the case where the samples are all from a from the case where one set of samples are from a and the other set of samples are from b.
   Hint: You may use the fact (without proving it) that distinguishing a fair coin from a coin of bias $\frac{1}{2}+\epsilon$ is known to require $\theta(1/\epsilon^2)$ flips.
   (b) Assume $\gamma > 1$. $\mathcal{A}$ is an algorithm that when given a distribution known to have at least $(\log n)/(\gamma^2)$ entropy will output a γ-approximation of the entropy of the distribution with probability at least 2/3. Show that $\mathcal{A}$ takes $\Omega(n^{1/(2\gamma^2)})$ samples from the distribution. (A proof sketch is sufficient – do not reprove claims from lecture or the last homework).
2. Let $\mathcal{S}$ be the class of distributions that are uniform over a subset of the domain D: $\mathcal{S} = \{U_S | S \subseteq D\}$.
   (a) What is the entropy of U_S in terms of $|S|$?
   (b) Show the following: Given $\gamma > 1$ and $p \in \mathcal{S}$ such that the support size of p is k. There is an algorithm which outputs ℓ such that $k/\gamma \leq \ell \leq \gamma k$ with probability at least 2/3 using $O(\gamma^2\sqrt{k}/(\gamma-1)^2)$ samples.
   (c) Show that this algorithm can be used to get a good additive estimation of the entropy for distributions that are in $\mathcal{S}$.
3. In the Run Length Encoding (RLE) compression scheme, the data is encoded as follows: each run, or a sequence of consecutive occurrences of the same character, is stored as a pair containing the character in the first location and the length of the run in the second location.[1] For example, the string 11111101000 would be stored as $(1,6)(0,1)(1,1)(0,3)$.

[1] Run-length encoding is used to compress black and white images, faxes, and other simple graphic images, such as icons and line drawings, which usually contain many long runs.

The *cost of the run-length encoding*, denoted by $C_{\mathrm{rle}}(w)$, is the sum over all runs of $\lceil \log(\ell+1) \rceil + \lceil \log |\Sigma| \rceil$.

Assume that the alphabet characters are all in the set $\{0,1\}$, i.e., that the alphabet Σ is of size 2.

(a) Give an algorithm that, given a parameter ϵ, outputs an ϵn-additive estimate[2] to $C_{\mathrm{rle}}(w)$ with high probability and makes $\mathrm{poly}(1/\epsilon, \log n)$ queries.

(b) Show that there is a distribution on inputs such any that any deterministic approximation algorithm for $C_{\mathrm{rle}}(w)$ making an expected number of queries that is $o(\frac{n}{A^2 \log n})$ must fail to output an A-multiplicative approximation with probability at least $1/3$. (Here the expectation in the number of queries is over the choice of an input from the distribution). (It's also ok to give a lower bound for deterministic algorithms by showing that for each algorithm there is an input that causes it to fail).

[2] For a function f, algorithm A outputs an ϵn-additive estimate if on any input x, $f(x) - \epsilon n \leq A(x) \leq f(x) + \epsilon n$.